

SENTIMENT ANALYSIS RESEARCH ON BENGALI LANGUAGE TEXTS

Dr. Vivek Kumar Singh

Asst. Prof., Dept., of Computer Science, Banaras Hindu University, Varanasi – 221 005, India.

ARTICLE INFO

Article History:

Received: 20 Sep 2015;

Received in revised form:

23 Sep 2015;

Accepted: 23 Sep 2015;

Published online: 30 Sep 2015.

Key words:

Bengali,

Opinion Mining,

Sentiment Analysis

Copyright © 2015 IJASRD. This is an open access article distributed under the Creative Common Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

This paper presents a scientometric survey of research work on sentiment analysis in Bengali language. All research papers published during 2008 to 2015 on sentiment analysis of Bengali texts, as indexed in Scopus, are identified and collected. The data is analyzed to find out the publication pattern and important works on sentiment analysis in Bengali language. The analytical results present a very informative account of research work on sentiment analysis in Bengali language.

INTRODUCTION

Sentiment analysis is a sophisticated language processing task, which infers emotions associated with a sentence or a document. With new transformed Word Wide Web, billions of users are creating content of different kinds on the Web. English continues to be the dominant language of Internet. However, during last few years user generated content in other languages is growing at a very rapid rate. Bengali is an important language with a substantially large number of speakers. This paper tries to survey the research work on sentiment analysis done in Bengali language. A scientometric methodology is used to collect research output data on sentiment analysis indexed in Scopus index. The research output is analyzed to identify main themes of research, levels of sentiment analysis work, types of datasets used etc.

SENTIMENT ANALYSIS RESEARCH IN BENGAI

Bengali is the second most spoken language in India with about 220 million native and about 300 million total speakers worldwide. Researchers have expressed their interest in Bengali text and there are many publications based on sentiment analysis with data resources from various Bengali corpuses have been published. This section will discuss the summary of few works in Bengali text.

Emotion Analysis: Das & Bandyopadhyay (2010a) presented two different approaches for identifying emotion holders from Bengali sentences. In this work, first

approach, the baseline model, is developed based on the combinations of various part-of-speech (POS) features extracted from the phrase-based similarities and the second approach, syntactic model, is based on the argument structure of the sentences with respect to the verbs. Bengali sentences extracted from Bengali blogs are used in this study. The performances of baseline model and syntactic model are then compared and concluded that syntactic model outperforms baseline model.

The work of Das & Bandyopadhyay (2010d) suggested a sentence level emotion tagging system based on the word level constituents. This work has been done with Bengali blog corpora and English news corpora. The training set for Bengali corpus was tagged manually with Ekman's six emotion tags. At first, the developed system semi-automatically carried out word level annotation. The baseline system and Condition Random Field (CRF) classifier, at word level for each emotion class assigns the class label to each word. The Conditional Random Field (CRF) based classifier used for word level emotion tagging outperformed the baseline for each emotion class. Sentence-level emotion scores for each emotion class are calculated as the average word level emotion scores based on the SentiWordNet. The emotion tag with the highest score is then assigned to the sentence, followed by a rule based post-processing technique for handling negative words.

In another study, (Das & Bandyopadhyay, 2009b) recommended an emotion analysis on Bengali blogs and News at word and sentence level. A set of six emotion tags, namely happy, sad, anger, fear, disgust and surprise have been selected for emotion detection task in this study. Emotional carrying words in a sentence are identified first and then Conditional Random Field (CRF) classifier assigns appropriate emotion tags to them. Weights of the emotion tags are then assigned by score based technique. A sense based scoring strategy identifies sentence level emotion scores for the six emotion tags based on the acquired word level emotion tags. Sentence level emotion tagging has been carried out based on the maximum obtained sentence level emotion scores. Similar approach done by Das & Bandyopadhyay (2009c) reported a word to sentence level emotion tagging on Bengali blogs. The proposed system consists of two phases. In the first phase, word level emotion classification is done by Conditional Random Field (CRF) classifier. The second phase assigns sentence level emotion tags based on the word level constituents using sense based scoring mechanism. The accuracy of the classifier is then measured by confusion matrix.

Another work done by Das & Bandyopadhyay (2011a) presented an approach to identify the emotions of the bloggers in Bengali blog documents. A simple rule based baseline system is designed to identify the emotional expression, holder and topic from the sentences. In addition to the baseline system, Support Vector Machine (SVM) based supervised framework is also employed to identify emotional expression, topic and holders. As the topics discussed by the bloggers at sentence level are not similar with the topics described at the document level, a semantic clustering approach has been adapted here for clustering the semantically related topic words present in the document.

In another work, Das & Bandyopadhyay (2011b) aimed to identify the emotional expressions at word, phrase, sentence and document level granularities along with their associated holders and topics. In this study, Conditional Random Field (CRF) and Support Vector Machine (SVM) are employed for word level emotional expression identification. Thereafter, sentence level emotion tagging is done through score base technique and the

phrase level emotional tagging is done based WordNet Affect lists and parsed dependency relations. Finally, the document level emotion tagging is carried out based on some combinations of heuristic features. This paper also proposed the techniques for holder and topic identifications. The proposed experiment is done with English and Bengali corpuses.

Incentive of the work suggested by Das et al. (2012) is to develop a blog based emotion analysis system for Bengali. This paper describes the identification, visualization and tracking of bloggers' emotions with respect to time from Bengali blog documents. The blog posts are collected from Bengali Web blog archive. In this work, the assignment of Ekman's six basic emotions to the bloggers' comments is carried out at word, sentence and paragraph level granularities using the Bengali WordNet Affect Lists.

In a book written by Das & Bandyopadhyay (2013), preparation of an emotion corpus and lexicon in Bengali is discussed. The emotion lexicon, termed Bengali WordNet Affect, has been developed from its equivalent version in English by traversing the steps of expansion, translation, and sense disambiguation. In addition to emotion lexicon, a Bengali blog corpus for emotion analysis has also been developed by manual annotators with detailed linguistic expressions such as emotional phrases, intensities, emotion holder, emotion topic and target span, and sentential emotion tags.

Theme Detection: The work proposed by Das & Bandyopadhyay (2009a) presented a theme detection technique for given set of documents. Sentiment analysis or opinion mining assumes that input documents are subjective but, it is always not true because some documents are objective. Moreover, it is essential to identify the features that make the documents subjective. Theme detection technique here is used to identify the subjectivity documents. The proposed rule-based technique assigns polarities for every sentence and finally accumulates the opinions to reach the discourse level subjectivity. Bengali news corpus and MPQA English corpus are used for experiments and then finally the performances are also compared.

Topic Based Opinion Summarization: Das & Bandyopadhyay (2010b) worked on topic based opinion summarization. Bengali news corpus has been developed from the archive of a leading Bengali newspaper available on the Web (<http://www.anandabazar.com/>) is used in this study. The proposed system identifies topic based sentiment information in each document, aggregates them and then presents summary of opinions. Topic based sentiment extraction task involves theme detection and theme clustering phases. Theme detection used Conditional Random Field (CRF) classifier to identify sentiment information from each document then the sentiments are aggregated by theme clustering algorithm, in this study k-means clustering algorithm. Finally, the Document Level Theme Relational Graph is finally used for candidate summary sentence selection by standard page rank algorithms used in Information Retrieval (IR). Similar approach proposed by Das & Bandyopadhyay (2010c) presented a theme network model for opinion summarization of Bengali news documents. In this paper, CRF classifier, k-means clustering algorithm and Theme Relational Graphs are used to detect theme based opinions and provide theme based opinion summaries.

Sentiment Detection: Azharul Hasan & Rahman (2014) proposed sentiment assessment of Bengali text using contextual valence analysis. The corpus consists of Bengali text is used in this study. The developed system in this paper first performs simple parsing to identify parts-of-speech then applies rules to assign contextual valence (polarity) to the

linguistic components in order to obtain sentence level sentiment valence. In another paper, Chowdhury & Chowdhury (2014) reported sentiment analysis in Bangla micro blog posts. Incentive of this work is to develop a system which can automatically extracts sentiments from the given Bangla text. The dataset is a collection of Bangla tweets downloaded by querying Twitter API v1.1. The proposed system uses semi-supervised bootstrapping approach and a rule-based classifier, to identify the sentiments and then classify them into predefined categories. Then, Support Vector Machine (SVM) and Maximum Entropy (ME) methods are also applied to compare the results.

A book written by Das & Gamback (2013) discusses various challenges for doing sentiment analysis on Bengali and also explains solution strategies for them. Another book written by Hasan et.al. (2013), presents a sentiment analyzer for recognizing sentence level sentiment or opinion about a subject from Bangla text. In this work, some phrase patterns are constructed and then their sentiment orientation is calculated. Thereafter, tags are added to the Bangla words to construct phrase pattern for positive and negative sentiment. Then, extracted phrase patterns in Bangla text are matched with the pre-defined patterns to calculate the sentiment orientation of each sentence.

The table 1 shows the types of datasets used in sentiment analysis research in Bengali. The table 2 presents a summary of important research works on this theme. The figure 1 shows the number of research papers on sentiment analysis in Bengali plotted year wise.

Table – 1: *Types of Datasets Used*

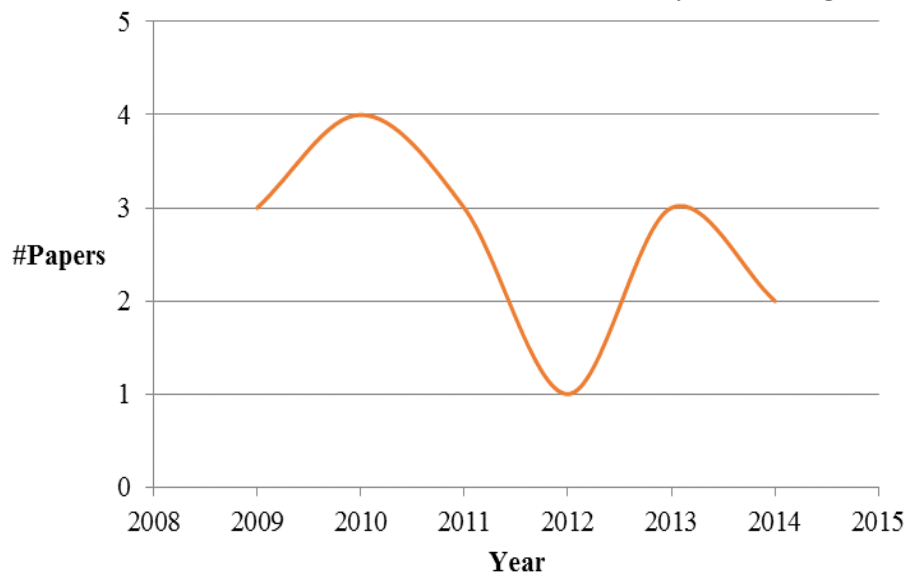
S. No.	Name	No. of Papers
1	Blogs	7
2	News	5
3	Papers	1
4	Bengali News	1
5	English	1
6	Bangla Tweets	1

Table – 2: *Sentiment Analysis Research Output in Bengali*

Topic	Method	Author	Dataset
Emotion analysis at word and sentence, document level, / identification of emotional expression, holder and topic / preparation of an emotion corpus and lexicon in Bengali Emotion	Conditional Random Field (CRF) classifier, score based technique, sense based scoring strategy	(Das and Bandyopadhyay 2009b)	Blogs, News
	Conditional Random Field (CRF) classifier, sense based scoring strategy	Das and Bandyopadhyay (2009c)	Blogs
	Baseline model and syntactic model	Das and Bandyopadhyay (2010a)	Blogs
	CRF and sentence level emotion score	Das and Bandyopadhyay (2010d)	Blogs, News papers
	rule based baseline system and Support Vector Machine (SVM) based supervised framework	Das and Bandyopadhyay (2011a)	Blogs

analysis	Word level- CRF and SVM; sentence level- score based technique; Phrase level- WordNet Affect lists and parsed dependency relations; document level- combinations of heuristic features	Das and Bandyopadhyay (2011b)	News, Blogs
	Bengali WordNet Affect Lists	Das et.al. (2012)	Blogs
		Das and Bandyopadhyay (2013)	
Theme detection	rule-based technique	Das and Bandyopadhyay (2009a)	Bengali News and English MPQA corpus
Topic based Opinion summarization	Theme detection-Conditional Random Field (CRF) classifier; sentiment aggregation-k means clustering; Summary generation - Theme Relational Graph	Das and Bandyopadhyay (2010b)	News
	Theme detection-Conditional Random Field (CRF) classifier; sentiment aggregation-k means clustering; Summary generation - Theme Relational Graph	Das and Bandyopadhyay (2010c)	News
Sentence level Sentiment Analysis/ Sentiment Detection	Pattern recognition	Hasan et.al. (2013)	
		Das and Gamback (2013)	
	contextual valence analysis	Azharul Hasan and Rahman (2014)	
	semi-supervised bootstrapping approach and a rule-based classifier	Chowdhury, S and Chowdhury, W (2014)	Bangla tweets

Figure – 1: Year wise Research Publication in Sentiment Analysis in Bengali



CONCLUSION

The paper surveys and summarizes the research output on sentiment analysis in Bengali language. Important research works on the theme are identified and summarized. Datasets used in different research works on the theme are summarized in table 1. The year-wise growth of research output is also plotted. The paper presents a very informative and useful account of sentiment analysis work in Bengali which is useful for people interested in sentiment analysis in Bengali.

REFERENCES

- [1] Chowdhury, S. & Chowdhury, W. (2014). Performing sentiment analysis in Bangla microblog posts. In Informatics, Electronics & Vision (ICIEV), 2014 International Conference on (pp. 1-6). IEEE.
- [2] Das D. and Bandyopadhyay S. (2010a). Finding emotion holder from Bengali blog texts -An unsupervised syntactic approach. PACLIC 24 - Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation.
- [3] Das, A. & Bandyopadhyay S. (2009a). Theme detection an exploration of opinion subjectivity. In Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on (pp. 1-6). IEEE.
- [4] Das A. & Bandyopadhyay S. (2010b). Topic-based Bengali opinion summarization. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 232-240). Association for Computational Linguistics.
- [5] Das, A. & Bandyopadhyay S. (2010c). Opinion Summarization in Bengali: A Theme Network Model. In Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 675-682). IEEE.
- [6] Das, A. & Gambäck B. (2013). যন্ত্র-না (Jantra-Na: Not-Machine) Can Only Feel যন্ত্রনা (Jantrana: Pain)!. Technical Challenges and Design Issues in Bangla Language Processing, 328.
- [7] Das, D. (2011b). Analysis and tracking of emotions in english and bengali texts: a computational approach. In Proceedings of the 20th international conference companion on World wide web (pp. 343-348). ACM.
- [8] Das, D. & Bandyopadhyay S. (2009b). Analyzing Emotion in Blog and News at Word and Sentence Level. In IICAI (pp. 1402-1414).
- [9] Das D. & Bandyopadhyay S. (2009c). Word to sentence level emotion tagging for Bengali blogs. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 149-152). Association for Computational Linguistics.
- [10] Das D. & Bandyopadhyay S. (2010d). Sentence level emotion tagging on blog and news corpora. Journal of Intelligent Systems, 19(2), 145-162.
- [11] Das D. & Bandyopadhyay S. (2011a). Emotions on Bengali blog texts: role of holder and topic. In Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on (pp. 587-592). IEEE.
- [12] Das D. & Bandyopadhyay, S. (2013). Building Language Resources for Emotion Analysis in Bengali. Technical Challenges and Design Issues in Bangla Language Processing, 346.
- [13] Das D., Roy S. & Bandyopadhyay S. (2012). Emotion tracking on blogs-a case study for bengali. In Advanced Research in Applied Artificial Intelligence (pp. 447-456). Springer Berlin Heidelberg.
- [14] Das D., Roy S. & Bandyopadhyay S. (2012). Emotion tracking on blogs-a case study for bengali. In Advanced Research in Applied Artificial Intelligence (pp. 447-456). Springer Berlin Heidelberg.
- [15] Hasan K. A., Islam M. S., Mashrur-E-Elahi G. M. & Izhar M. N. (2013). Sentiment Recognition from Bangla Text. Technical Challenges and Design Issues in Bangla Language Processing, 315.